Reliability of the BI-RADS Final Assessment Categories and Management Recommendations in a Telemammography Context

Antonio J. Salazar, PhD[a], Javier A. Romero, MD, MSc[b], Oscar A. Bernal, MD, PhD[c], Angela P. Moreno, MD, MSc[b], Sofía C. Velasco, MD, MSc[b]

Abstract

Purpose: The aim of this study was to evaluate the intradevice and interdevice reliability of four alternatives for telemammography—computed radiography, printed film, a film digitizer, and a digital camera—in terms of interpretation agreement when using the BI-RADS® lexicon.

Methods: The ethics committee of the authors' institution approved this retrospective study. A factorial design with repeated measures with 1,960 interpretations was used (70 patients, seven radiologists, and four devices). Reliability was evaluated using the κ coefficient for intradevice and interdevice agreement on malignancy classification and on BI-RADS final assessment category.

Results: Agreement on malignancy classification was higher than agreement for BI-RADS final assessment category. Interdevice agreement on malignancy classification between the film digitizer and computed radiography was ranked as almost perfect ($P < .001$), whereas interdevice agreement for the other alternatives was ranked as substantial ($P < .001$), with observed agreement ranging from 85% to 91% and κ values ranging from 0.70 to 0.81. Interdevice agreement on BI-RADS final assessment category was ranked as substantial or moderate ($P < .001$), with observed agreement ranging from 64% to 77% and κ values ranging from 0.52 to 0.69. Interdevice agreement was higher than intradevice agreement.

Conclusions: The results of this study show very high interdevice agreement, especially for management recommendations derived from malignancy classification, which is one of the most important outcomes in screening programs. This study provides evidence to suggest the interchangeability of the devices evaluated, thereby enabling the provision of low-cost medical imaging services to underserved populations.

Key Words: BI-RADS, breast cancer, film digitizer, mammography, reliability, telemammography

J Am Coll Radiol 2017;14:686-692.

INTRODUCTION

Variability in radiologists' interpretations may reduce the accuracy of mammography in the early detection of breast cancer [1]. To standardize the reporting of findings in different imaging modalities, the ACR developed the BI-RADS® atlas [2]. Several studies have evaluated the accuracy of this system compared with traditional mammography [3,4]. Other studies have evaluated the accuracy of mammography, ultrasound, and physical examination, compared with biopsy findings, when using the BI-RADS lexicon [5,6].

Assessments regarding the BI-RADS atlas usually concern feature analysis (eg, breast density, lesion type, mass borders, mass density, mass shape, microcalcification morphology, microcalcification distribution [3]), as well as assessments of management recommendations (eg, routine mammographic screening, short-interval follow-up, tissue diagnosis) [3-5]. According to these studies, results obtained using the BI-RADS categories have been found to be useful in differentiating between benign and malignant breast lesions [7].

[a]Electrophysiology and Telemedicine Laboratory, University of Los Andes, Bogotá, Colombia.

[b]Department of Diagnostic Imaging, Fundación Santa Fe de Bogotá University Hospital, Bogotá, Colombia.

[c]School of Government, University of Los Andes, Bogotá, Colombia.

Corresponding author and reprints: Antonio J. Salazar, PhD, Electrophysiology and Telemedicine Laboratory, University of Los Andes, Carrera 1 Este No. 19A-40, Bogotá, Colombia; e-mail: ant-sala@uniandes.edu.co.

The authors have no conflicts of interest related to the material discussed in this article.

1546-1440/16 ■ http://dx.doi.org/10.1016/j.jacr.2016.08.004

In underserved areas, telemedicine, using digital images, may provide a cost-effective solution for screening mammography programs. Previous studies have reported no significant differences between screen-film mammography and digital modalities, such as computed radiography (CR) and full-field digital mammography [8,9]. Nevertheless, these technologies are still unaffordable for the vulnerable populations of Colombia, especially in the Amazonian jungles where only conventional screen-film mammography is available, if any radiologic services exist at all [10]. In previous studies, we evaluated the validity of low-cost telemammography configurations, in terms of sensitivity, specificity, and receiver operating characteristic curves [11]. In this study, the aim was to assess reliability among different solutions for telemammography, such as film digitizers, digital cameras, printed film, and CR, in terms of interpretation agreement, over interpretation results based on the BI-RADS lexicon.

METHODS

The ethics committee of our institution approved this retrospective study, and informed consent was not required. This study applied a design with repeated measures, using 70 patients, seven radiologists, the reference images (ie, CR) and three derived images, for a total of 1,960 readings.

To perform validity assessments, sensitivity, specificity, and positive predictive value are usually evaluated. In contrast, in the context of this study, reliability was the reproducibility or agreement in measurements of the variables for each case when rated by different observers (ie, intradevice reliability) or when rated by each observer using different treatments (ie, interdevice reliability).

The Reference Standard

The actual state of the mammograms enabled us to determine the distribution of the sample. The standard for positive cases was a malignant lesion confirmed by biopsy within 2 years of the initial mammographic screening [8,9,12]. Negative cases were defined as those without any lesions confirmed by biopsy or those with normal results on follow-up mammography for 2 years. Two radiologists with more than 10 years of experience in reading mammograms, with access to the clinical histories of the patients, established the reference standard.

Study Sample and Readers

At most rural health centers in our country, there are no mammographic services [10]. As such, there are no mammograms available for use in a retrospective study. For these reasons, this study was undertaken using computed radiographic screening mammograms from our hospital, which sees many patients from such underserved areas of our country.

Random screening mammograms from asymptomatic patients who attended mammographic screening at the Fundación Santa Fe de Bogotá University Hospital, performed over 2 years, were included in the study sample. Of these patients, 50% had management recommendations for tissue diagnosis and 50% for follow-up. To be included, each case was required to include the four standard mammographic views. Cases of tomosynthesis or large lesions were excluded.

The sample size was determined in our previous studies to be 70 cases, with an approximately 1:1 ratio of malignant and benign cases. The patients ranged in age from 41 to 84 years, with a mean age of 62.1 $\pm$ 11.5 years. There were 57 cases with calcifications, 26 with masses, 35 with asymmetries, and 11 with architectural distortions and associated features. Four patients with prostheses were also included in the sample.

The distribution of cases according to BI-RADS final assessment category was as follows: 18 in category 2, 19 in category 3, 6 in category 4A, 14 in category 4B, 3 in category 4C, and 10 in category 5. In terms of malignancy, there were 33 patients with malignant lesions and 37 patients with benign lesions or normal results. Detailed classification of the cases in the sample, and their distribution according to the BI-RADS final assessment categories are presented in supplemental tables (S1 and S2).

Seven radiologists from Fundación Santa Fe de Bogotá University Hospital (including four with more than 10 years of experience in mammography who were well trained in the BI-RADS lexicon and three radiologists with more than 1 year of experience who were trained in the BI-RADS lexicon for the purposes of this study) served as the observers.

Observed Variables by Radiologists

Data collection was performed using a database and a digital form that was integrated into the image-viewing software. At each interpretation, the radiologist selected the level of confidence in the presence of masses, calcifications, asymmetries, and architectural distortions and associated features. The radiologists were asked to classify the lesion features, such as mass borders, mass density, mass shape, microcalcification morphology, microcalcification distribution, asymmetric density, and architectural distortion. Additionally, the radiologist classified

the breast composition. Finally, as a conclusion to the interpretation process, the radiologist selected a BI-RADS final assessment category. This variable was used at the present assessment as the variable to evaluate reliability.

Criteria to Establish a BI-RADS Final Assessment Category

To improve the consistency of mammographic interpretations at our hospital and to reduce radiologist variability in final assessments and in management recommendations, radiologists are trained in criteria to establish the BI-RADS final assessment category, according to specific mammographic features and their positive predictive value for malignancy (eg, the highest positive predictive value of malignancy included masses with spiculated margins, irregular shape, calcifications with linear morphology, or segmental distribution) [7]. The specific mammographic features associated with the BI-RADS final assessment categories defined by the criteria were as follows: no findings to report classified as category 1; global asymmetry (not palpable lesions), benign calcifications (calcified fibroadenomas, skin calcifications), and metallic foreign bodies classified as category 2; asymmetry, focal asymmetry, solitary group of punctate calcifications, and noncalcified circumscribed solid mass (not palpable lesions) classified as category 3; global asymmetry (palpable lesions), coarse heterogeneous calcification, and well-defined mass (not palpable) classified as category 4A; developing asymmetry, amorphous calcification, microlobulated mass, and obscured edge mass classified as category 4B; fine pleomorphic calcifications and poorly defined mass classified as category 4C; and pleomorphic ductal pattern and spiculated mass classified as category 5 (Table S3).

Algorithm for Automated Selection of the BI-RADS Assessment

The criteria taught to the radiologists for establishing the BI-RADS final assessment category were implemented in the database during the data analysis process. An algorithm to determine the BI-RADS final assessment category, according to the radiologists' lesion findings and the established criteria, was then applied. This category value was compared with the actual value selected by the radiologist in each interpretation, and an agreement evaluation between these two values was performed. This comparison enabled us to evaluate how effectively the radiologists were trained in the criteria to establish the BI-RADS final assessment category.

Generation and Digitization of Mammograms

The original screening mammograms consisted of computed radiographic images stored in the PACS at our hospital. Computed radiographic images were acquired using an Agfa CR 85-X (Agfa HealthCare NV, Mortsel, Belgium), with a resolution of 50 μm/pixel and a 14-bit grayscale and 3,560 × 4,640 pixel matrix. These computed radiographic images were printed on 18 × 24 cm film with a digital Agfa Drystar 5503 printing system (Agfa HealthCare NV) with resolution of 50 μm/pixel and 14-bit contrast. Data that could be used to identify patients were not printed. Next, the films were digitized using the following capture devices: (1) a specialized digitizer, iCR 612SL (iCR Company, Torrance, California), that had a maximum spatial resolution of 875 dpi, a pixel spot of 29 μm, 16 bits/pixel, an optical density of 3.6, and a cost of $15,000, and (2) a Lumix DMC-FZ28 digital camera (Panasonic Corporation, Secaucus, New Jersey) with 10-megapixel resolution, a focal length of 4.8 to 86.4 mm, a 1/2.33-inch charge-coupled device, ISO 100-6400, and a cost of $450.

For each patient, the following case studies were obtained: (1) the printed film and three digital images, including (2) images from CR (3,560 × 4,640 pixel matrix and 14-bit grayscale), (3) images digitized with the iCR digitizer (2,436 × 3,636 pixel matrix and 8-bit grayscale), and (4) images digitized with the Lumix camera (2,538 × 3,463 pixel matrix and 8-bit grayscale). A DICOM-compliant software package that was developed at our institution [13] was used to scan, store, and display the cases.

Data Analysis

We evaluated agreement using the BI-RADS final assessment category as follows: (1) agreement on single BI-RADS final assessment category and (2) agreement on malignancy classification using a dichotomized variable with a value of 0 for negative findings (BI-RADS categories 2 and 3) and 1 for positive findings (BI-RADS categories 4A, 4B, 4C, and 5).

For the variables malignancy classification and individual BI-RADS final assessment category, we evaluated both intradevice agreement (ie, agreement between radiologists when interpreting using a single device) and interdevice agreement (ie, agreement for radiologists when interpreting the same patient using images from two different devices). As a result, we evaluated the following variables: (1) intradevice agreement on malignancy classification, (2) interdevice agreement on malignancy

classification, (3) intradevice agreement on BI-RADS final assessment category, and (4) interdevice agreements on BI-RADS final assessment category. These four variables were evaluated with the κ coefficient as a measure of agreement. The κ coefficients were ranked as defined by Landis and Koch [14] as follows: perfect, $\kappa = 1$; almost perfect, $\kappa = 1$ to 0.8; substantial, $\kappa = 0.8$ to 0.6; moderate, $\kappa = 0.6$ to 0.4; fair, $\kappa = 0.4$ to 0.2; slight, $\kappa = 0.2$ to 0; and Poor, $\kappa < 0$. For these calculations, IBM SPSS Statistics 19 (IBM, Armonk, New York) was used.

Procedure

All cases were read by each radiologist using the following viewing methods: the film in a lightbox and three viewings on a medical display for digital cases of CR, the iCR digitizer, and the Lumix camera. A DICOM-compliant 3-megapixel MD213MG (NEC Display Solutions, Tokyo, Japan) medical-grade grayscale display was used as the display monitor. The software provides image manipulation tools to adjust the window and level and histogram tools (eg, average optical density, histogram equalization, and full-scale histogram stretch). These tools could be combined with the zoom and the magnifying glass. The radiologists were blinded by the software to the patient and examination information and to the capture device producing the images. Pairs of patients and capture devices were presented at random by the software. The readings were performed over the course of 10 months in 2- or 4-hour sessions by each radiologist, with no time limitations for each reading.

RESULTS

Intradevice Agreements

The intradevice agreement observed in our study is presented in Table 1 and is separated into three groups: (1) intradevice agreement on malignancy classification; (2) intradevice agreement on BI-RADS final assessment category without grouping categories 4A, 4B, and 4C; and (3) intradevice agreement on BI-RADS final assessment category grouping categories 4A, 4B, and 4C. Global and individual category agreement is presented for this group.

Agreement between radiologists on malignancy classification was higher than agreement for BI-RADS final assessment category. Intradevice agreement on malignancy classification for all devices showed κ values that were ranked as moderate ($P < .001$) for digital modalities (ie, CR, iCR, and Lumix) and substantial for film ($P < .001$). In contrast, the global intradevice agreement on BI-RADS final assessment category by device, without grouping categories 4A, 4B, and 4C, was ranked as fair for all devices, as follows: $\kappa = 0.33$ for CR, $\kappa = 0.38$ for iCR, $\kappa = 0.31$ for Lumix, and $\kappa = 0.38$ for film, with 95% confidence intervals with lower bounds ranked as fair and upper bounds as fair for Lumix and moderate for CR, iCR, and film. The global intradevice agreement on BI-RADS final assessment category by device, grouping categories 4A, 4B, and 4C, was ranked as fair ($P < .001$), with higher κ values for film and iCR, whereas the computed radiographic original images produced a κ value of 0.40. The 95% confidence intervals when grouping categories 4A, 4B, and 4C were higher than when they were not grouped, with all upper bounds agreement ranked as moderate. With regard to the κ values for individual categories, category 3 had the lowest agreement for all devices.

Interdevice Agreement

Interdevice agreement on malignancy classification and on BI-RADS final assessment category by device is presented in Table 2. Agreement on malignancy classification between iCR and CR was ranked as almost perfect ($P < .001$), whereas interdevice agreement for the other alternatives was ranked as substantial ($P < .001$), with observed agreement ranging from 85.5% to 91.0% and κ values ranging from 0.70 to 0.81.

Lower κ values than those for agreement on malignancy classification were observed for interdevice agreement on BI-RADS final assessment category, as follows: agreement among paired devices were ranked as substantial or moderate ($P < .001$), with observed agreement ranging from 63.9% to 76.5% and κ values ranging from 0.52 to 0.69.

Agreement Between Radiologists and the Criteria to Establish a BI-RADS Final Assessment Category

According to the radiologists' lesion findings for each interpretation, the software calculated the corresponding category, and the agreement of this value with the actual value selected by the radiologists was evaluated. Analysis of intradevice agreement on single BI-RADS final assessment categories showed intradevice observed agreement greater than 90%, with high κ values, ranging from 0.89 to 0.90, and all were ranked as almost perfect. Analysis of agreement on malignancy classification showed intradevice observed agreement greater than

Table 1. Intradevice agreement by device

Device	Category	κ*	SE	95% CI LB	95% CI UB	z	P	Agreement†
On malignancy classification								
CR	Global	0.55	0.058	0.44	0.66	21.1	<.001	Moderate
iCR	Global	0.54	0.062	0.42	0.66	20.8	<.001	Moderate
Lumix	Global	0.49	0.059	0.37	0.60	18.6	<.001	Moderate
Film	Global	0.63	0.056	0.52	0.74	24.3	<.001	Substantial
On BI-RADS final assessment category without grouping categories 4A, 4B, and 4C								
CR	Global	0.33	0.039	0.26	0.41	24.4	<.001	Fair
iCR	Global	0.38	0.044	0.29	0.46	27.0	<.001	Fair
Lumix	Global	0.31	0.040	0.23	0.39	22.0	<.001	Fair
Film	Global	0.38	0.038	0.30	0.45	27.4	<.001	Fair
On BI-RADS final assessment category grouping categories 4A, 4B, and 4C								
CR	2	0.44	0.061	0.32	0.56	17.0	<.001	Moderate
	3	0.23	0.047	0.14	0.33	9.0	<.001	Fair
	4	0.45	0.059	0.33	0.57	17.2	<.001	Moderate
	5	0.66	0.183	0.30	1.02	25.3	<.001	Substantial
	Global	0.40	0.045	0.31	0.49	23.7	<.001	Fair
iCR	2	0.48	0.060	0.36	0.60	18.5	<.001	Moderate
	3	0.35	0.058	0.23	0.46	13.3	<.001	Fair
	4	0.45	0.065	0.32	0.58	17.3	<.001	Moderate
	5	0.68	0.195	0.29	1.06	26.0	<.001	Substantial
	Global	0.44	0.048	0.35	0.54	26.1	<.001	Moderate
Lumix	2	0.38	0.053	0.28	0.49	14.7	<.001	Fair
	3	0.20	0.043	0.12	0.29	7.8	<.001	Fair
	4	0.40	0.057	0.29	0.52	15.5	<.001	Moderate
	5	0.66	0.208	0.25	1.07	25.5	<.001	Substantial
	Global	0.35	0.043	0.27	0.44	20.6	<.001	Fair
Film	2	0.52	0.063	0.39	0.64	19.9	<.001	Moderate
	3	0.35	0.056	0.24	0.46	13.6	<.001	Fair
	4	0.52	0.059	0.41	0.64	20.1	<.001	Moderate
	5	0.60	0.170	0.26	0.93	22.8	<.001	Moderate
	Global	0.47	0.044	0.39	0.56	28.0	<.001	Moderate

Note: CI = confidence interval; CR = computed radiography; LB = lower bound; UB = upper bound.
*Each agreement level was calculated from 490 readings (70 cases by seven radiologists).
†As defined by Landis and Koch [14].

96%, with high κ values, ranging from 0.93 to 0.96, and all were ranked as almost perfect (Table S4).

DISCUSSION

Intradevice agreement on BI-RADS assessment categories was ranked as fair for all devices, both including and excluding categories 4A, 4B, and 4C. These results in our study are in accordance with the results of some other studies using conventional mammography (with no digitizing process), in which fair agreement was found [3]. Although the κ values were low for the BI-RADS final assessments categories, even when categories 4A, 4B, and 4C were grouped, the observed agreement on malignancy classification was high (ranked as moderate or substantial). This is in line with the results of other studies evaluating malignancy classification, cancer detection, or management recommendation, in which the following final assessment categories were grouped: follow-up (for categories 1, 2, and 3 combined) and biopsy (for categories 4A, 4B, 4C, and 5 combined) [3,5,15].

For interdevice agreement on malignancy classification, very high agreement was noted (ranked substantial or almost perfect). The interdevice agreement on BI-RADS final assessments category was lower than agreement on malignancy classification, which corresponds again to the previously reported result of grouping categories. Nevertheless, high κ values were observed and were ranked as substantial or moderate.

Intradevice agreement on the individual BI-RADS assessment categories (2, 3, 4, and 5) showed lower

Table 2. Interdevice agreement

Devices	OA (%)	EA (%)	κ*	SE(0)	$z = \kappa/SE(0)$	P	Agreement†
On malignancy classification							
Lumix vs iCR	88.8	53.2	0.76	0.045	16.8	<.001	Substantial
Lumix vs CR	90.4	52.7	0.80	0.045	17.7	<.001	Substantial
Lumix vs Film	85.5	52.4	0.70	0.045	15.4	<.001	Substantial
iCR vs CR	91.0	53.2	0.81	0.045	17.9	<.001	Almost Perfect
ICR vs Film	87.4	52.8	0.73	0.045	16.3	<.001	Substantial
CR vs Film	86.1	52.3	0.71	0.045	15.7	<.001	Substantial
On BI-RADS final assessment category							
Lumix vs ICR	76.5	24.4	0.69	0.024	28.6	<.001	Substantial
Lumix vs CR	76.3	23.8	0.69	0.024	28.9	<.001	Substantial
Lumix vs Film	63.9	24.0	0.52	0.024	21.9	<.001	Moderate
iCR vs CR	75.9	24.0	0.68	0.024	28.5	<.001	Substantial
iCR vs Film	67.1	24.0	0.57	0.024	23.7	<.001	Moderate
CR vs Film	64.1	23.5	0.53	0.024	22.4	<.001	Moderate

Note: CR = computed radiography; EA = expected agreement; OA = observed agreement; SE(0) = κ standard error (H_0: κ = 0).
*Each agreement level was calculated from 980 readings (70 cases by seven radiologists by two devices).
†As defined by Landis and Koch [14].

agreement for category 3 (probably benign). Even though several studies have reported that the BI-RADS classification is an accurate method for differentiating between benign and malignant lesions [4,5,15], BI-RADS final assessment category 3 is the limit between biopsy or mammographic follow-up and thus is the limit for conservative management to avoid unnecessary biopsies. As such, it is a good predictor of benignity. On the other hand, selecting category 3 assessments directly from screening examinations has been shown to result in the following adverse outcomes: (1) unnecessary follow-up of many lesions that could have been promptly assessed as benign and (2) delayed diagnosis of a small number of cancers that otherwise may have been smaller in size and less likely to be advanced in stage [2]. In the process of selecting category 3, it is important to evaluate palpable or not palpable lesions, as according to Harvey et al [16], short-term follow-up is a reasonable alternative to biopsy of palpable breast lesions with benign imaging features. Additionally, according to Graf et al [17], palpable noncalcified solid breast masses with benign morphology at mammography may be managed similarly to nonpalpable BI-RADS category 3 lesions, with short-term follow-up. Nevertheless, it is not prudent to render a category 3 assessment when a finding that otherwise meets "probably benign" imaging criteria is either new or has increased in size or extent [2,18,19]. One limitation of our study was that the radiologists only had access to images and were blinded to patient histories. Consequently, it was not possible for them to determine if a finding was a new finding or a lesion that had increased in size. Therefore, category 3 could not be easily selected by the radiologists in this study, thereby producing low agreement rates for this category.

CONCLUSIONS

To the best of our knowledge, evaluations and comparisons among low-cost devices for telemammography, on the basis of the BI-RADS final assessment categories and management recommendations, have not been reported thus far. The results of our study show very high interdevice agreement, especially for management recommendations. In other studies, the focus was placed on assessing agreement with regard to the description of the lesion features or the accuracy of the BI-RADS lexicon per se. However, in this study, the focus was to assess the concordance of individual devices, and in paired devices, to evaluate reliability for low-cost telemammographic solutions. This study provides evidence of high agreement in paired device evaluations on malignancy classification (associated with clinical management recommendations), suggesting the possibility of using the low-cost devices evaluated in our study in telemammographic screening programs, thereby providing high-quality medical imaging services to underserved populations.

TAKE-HOME POINTS

- The criteria and training to establish a BI-RADS final assessment category improved the consistency in mammographic interpretations.

- The results of our study showed very high inter-device agreement on management recommendations among the evaluated devices (film, CR, an iCR digitizer, and a Lumix digital camera) on the basis of the BI-RADS final assessment categories.
- Reliability agreement in paired device evaluations was confirmed.
- Our results suggest the possibility of using the less expensive devices evaluated in this study in telemammographic screening programs.
- High-quality screening telemammographic services may be provided to underserved populations at low cost.

ACKNOWLEDGMENTS

We thank the radiologists who carried out the interpretations. We thank our institutions and the National Department of Science, Technology and Innovation for funding this study (grant 1204-545-31353).

ADDITIONAL RESOURCES

Additional resources can be found online at: http://dx.doi.org/10.1016/j.jacr.2016.08.004.

REFERENCES

1. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331:1493-9.
2. Sickles EA, D'Orsi CJ, Bassett LW. ACR BI-RADS® mammography. Reston, Virginia: American College of Radiology; 2013.
3. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. AJR Am J Roentgenol 2000;174:1769-77.
4. McKay C, Hart C, Erbacher G. Objectivity and accuracy of mammogram interpretation using the BI-RADS final assessment categories in 40- to 49-year-old women. J Am Osteopath Assoc 2000;100:615-20.
5. Lorenzen J, Wedel AK, Lisboa BW, Löning T, Adam G. Diagnostic mammography and sonography: concordance of the breast imaging reporting assessments and final clinical outcome. Fortschr Röntgenstr 2005;177:1545-51.
6. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. Radiology 2002;225:165-75.
7. Liberman L, Abramson AF, Squires FB, et al. The Breast Imaging Reporting and Data System: positive predictive value of mammographic features and final assessment categories. AJR Am J Roentgenol 1998;171:35-40.
8. Gitlin J, Narayan A, Mitchell C, et al. A comparative study of conventional mammography film interpretations with soft copy readings of the same examinations. J Digit Imaging 2007;20:42-52.
9. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med 2005;353:1773-83.
10. Velasco S, Bernal O, Salazar A, et al. Availability of mammography services in Colombia. Rev Colomb Cancerol 2014;18:101-8.
11. Salazar A, Romero J, Bernal O, et al. Evaluation of low-cost telemammography screening configurations: A Comparison with film-screen readings in vulnerable areas.
12. Lewin JM, D'Orsi CJ, Hendrick RE, et al. Clinical comparison of full-field digital mammography and screen-film mammography for detection of breast cancer. AJR Am J Roentgenol 2002;179:671-7.
13. Salazar AJ, Aguirre DA, Ocampo J, Camacho JC, Díaz XA. DICOM gray-scale standard display function: Clinical diagnostic accuracy of chest radiography in medical-grade gray-scale and consumer-grade color displays. AJR 2014;202:1272-80.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
15. Orel SG, Kay N, Reynolds C, Sullivan DC. BI-RADS categorization as a predictor of malignancy. Radiology 1999;211:845-50.
16. Harvey JA, Nicholson BT, LoRusso AP, Cohen MA, Bovbjerg VE. Short-term follow-up of palpable breast lesions with benign imaging features: evaluation of 375 lesions in 320 women. AJR Am J Roentgenol 2009;193:1723-30.
17. Graf O, Helbich TH, Fuchsjaeger MH, et al. Follow-up of palpable circumscribed noncalcified solid breast masses at mammography and US: can biopsy be averted? Radiology 2004;233:850-6.
18. Varas X, Leborgne JH, Leborgne F, et al. Revisiting the mammographic follow-up of BI-RADS category 3 lesions. AJR Am J Roentgenol 2002;179:691-5.
19. Vizcaíno I, Gadea L, Andreo L, et al. Short-term follow-up results in 795 nonpalpable probably benign lesions detected at screening mammography. Radiology 2001;219:475-83.

Table S1. Detailed classification of the cases in the sample

Condition	Classification*	Cases
Masses	Well-defined mass	7
	Obscured edge mass	10
	Poorly defined mass	4
	Spiculated mass	5
Calcifications	Benign calcifications	33
	Solitary group of punctate calcifications	4
	Coarse heterogeneous calcification	8
	Amorphous calcification	7
	Fine pleomorphic calcifications	4
	Pleomorphic ductal pattern	1
Architectural distortions and associated features		11
Asymmetries	Asymmetry	23
	Focal asymmetry	12

*Classification according to the ACR [2].

Table S2. Distribution of cases in the sample, according to BI-RADS final assessment category

BI-RADS Final Assessment Category*	Cases
2 = Benign	18
3 = Probably benign	19
4A = Low suspicion for malignancy	6
4B = Moderate suspicion for malignancy	14
4C = High suspicion for malignancy	3
5 = Highly suggestive of malignancy	10
Total	70

*Classification according to the ACR [2].

Table S3. Specific mammographic features associated with the BI-RADS final assessment categories, according to their PPV

BI-RADS Final Assessment Category	Specific Mammographic Features	PPV
1 = Normal	No findings to report	
2 = Benign	Global asymmetry (not palpable lesions)	0%
	Benign calcifications (calcified fibroadenomas, skin calcifications)	
	Metallic foreign bodies	
3 = Probably benign	Asymmetry	1.8%
	Focal asymmetry	0.67%
	Solitary group of punctate calcifications	0.67%
	Noncalcified circumscribed solid mass (not palpable lesions)	2.0%
4A = Low suspicion	Global asymmetry (palpable lesions)	7.5%
	Coarse heterogeneous calcification	7.0%
	Well-defined mass (not palpable)	2%-10%
4B = Intermediate suspicion	Developing asymmetry	12.1%
	Amorphous calcification	12%-26%
	Microlobulated mass	17%-50%
	Obscured edge mass	13%-33%
4C = Moderate suspicion	Fine pleomorphic calcifications	79%
	Poorly defined mass	51%-94%
5 = Highly suggestive of malignancy	Pleomorphic ductal pattern	70%-100%
	Spiculated mass	64%-100%

Note: PPV = positive predictive value.

Table S4. Agreement between radiologists and automated selection on BI-RADS final assessment category by device

Devices	OA (%)	EA (%)	κ*	SE(0)	$z = \kappa/SE(0)$	*P*	Agreement†
On BI-RADS final assessment							
CR	91.4	23.9	0.89	0.024	37.0	<.001	Almost perfect
iCR	92.7	25.1	0.90	0.025	36.9	<.001	Almost perfect
Lumix	91.2	24.7	0.88	0.024	36.3	<.001	Almost perfect
Film	90.8	23.9	0.88	0.024	36.9	<.001	Almost perfect
On malignancy classification							
CR	97.8	52.9	0.95	0.045	21.1	<.001	Almost perfect
iCR	97.6	54.3	0.95	0.045	21.0	<.001	Almost perfect
Lumix	96.7	52.9	0.93	0.045	20.6	<.001	Almost perfect
Film	98.2	52.3	0.96	0.045	21.3	<.001	Almost perfect

Note: EA = expected agreement; OA = observed agreement; SE(0) = κ standard error (H_0: κ = 0).
*Each agreement level was calculated from 490 readings (70 cases by seven radiologists).
†As defined by Landis and Koch [14].